# crossentropy_method

July 21, 2017

## 1 References

This notebook is heavily based on the excellent "Practical RL" course from the Yandex School of Data Analysis https://github.com/yandexdataschool/Practical_RL/

## 2 Crossentropy method

This notebook will teach you to solve reinforcement learning with crossentropy method.

```
In [1]: #XVFB will be launched if you run on a server
        import os
        if type(os.environ.get("DISPLAY")) is not str or len(os.environ.get("DISPLAY"))==0:
            !bash ../xvfb start
            %env DISPLAY=:1
        import matplotlib.pylab as plt
        %matplotlib inline
```

```
In [2]: import gym
        import numpy as np, pandas as pd

        env = gym.make("Taxi-v2")
        env.reset()
        env.render()
```

```
[2017-07-21 10:26:21,650] Making new env: Taxi-v2


+---------+
|R: | : :G|
| : : : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
In [3]: n_states = env.observation_space.n
        n_actions = env.action_space.n

        print("n_states=%i, n_actions=%i"%(n_states,n_actions))

n_states=500, n_actions=6
```

## 3 Create stochastic policy

This time our policy should be a probability distribution.

`policy[s,a] = P(take action a | in state s)`

Since we still use integer state and action representations, you can use a 2-dimensional array to represent the policy.

Please initialize policy **uniformly**, that is, probabililities of all actions should be equal.

```
In [4]: policy = np.ones((n_states, n_actions)) / n_actions
```

```
In [5]: assert type(policy) in (np.ndarray,np.matrix)
        assert np.allclose(policy,1./n_actions)
        assert np.allclose(np.sum(policy,axis=1), 1)
```

## 4 Play the game

Just like before, but we also record all states and actions we took.

```
In [8]: def generate_session(policy, t_max=10**4):
            """
            Play game until end or for t_max ticks.
            returns: list of states, list of actions and sum of rewards
            """
            states,actions = [],[]
            total_reward = 0.

            s = env.reset()

            for t in range(t_max):

                a = np.random.choice(n_actions, 1, p=policy[s, :])

                new_s,r,done,info = env.step(a[0])

                states.append(s)
                actions.append(a)
                total_reward += r

                s = new_s
                if done:
                    break
            return states,actions,total_reward
```

```
In [10]: s,a,r = generate_session(policy)
         assert type(s) == type(a) == list
         assert len(s) == len(a)
         assert type(r) is float
```

## 5 Training loop

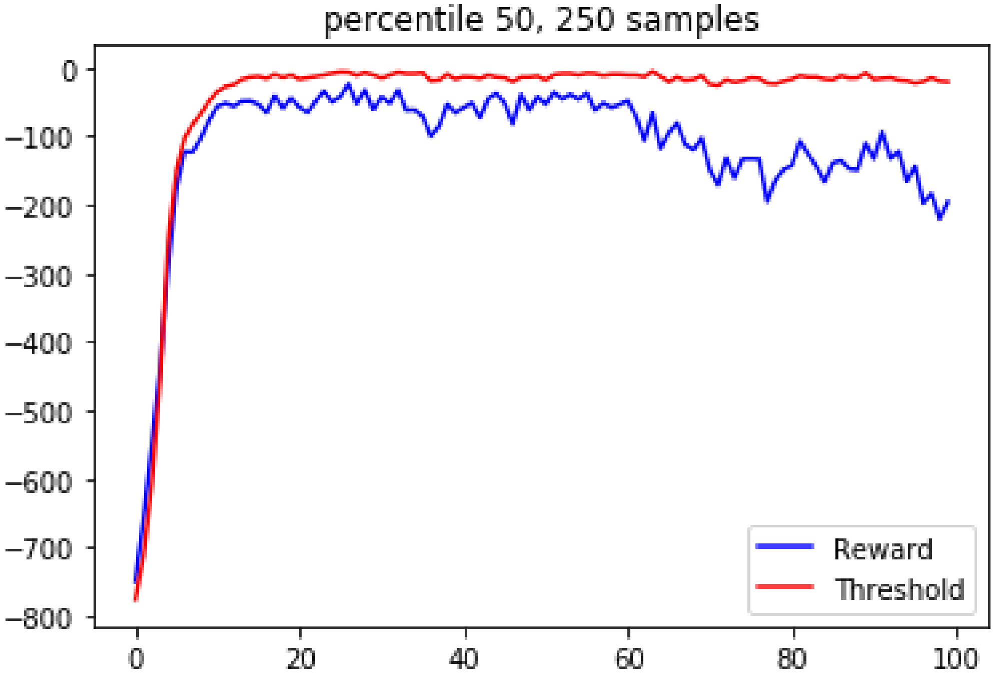
Generate sessions, select N best and fit to those.

```
In [13]: def run(policy, n_samples=250, percentile=50, smoothing=.1):
             step_rewards = []
             step_thresholds = []

             for i in range(100):

                 sessions = [generate_session(policy) for k in range(n_samples)]

                 batch_states,batch_actions,batch_rewards = map(np.array,zip(*sessions))

                 #batch_states: a list of lists of states in each session
                 #batch_actions: a list of lists of actions in each session
                 #batch_rewards: a list of floats - total rewards at each session

                 threshold = np.percentile(batch_rewards, percentile)

                 elite_states = [batch_states[j] for j, r in enumerate(batch_rewards) if r > t
                 elite_actions  = [batch_actions[j] for j, r in enumerate(batch_rewards) if r

                 elite_states, elite_actions = map(np.concatenate,[elite_states,elite_actions]
                 #hint on task above: use np.percentile and numpy-style indexing

                 #count actions from elite states
                 elite_counts = np.zeros_like(policy)+smoothing

                 for s, a in zip(elite_states, elite_actions):
                     elite_counts[s, a] += 1

                 policy = elite_counts / elite_counts.sum(axis=1).reshape(elite_counts.shape[0

                 print("mean reward = %.5f\tthreshold = %.1f"%(np.mean(batch_rewards),threshol
                 step_rewards.append(np.mean(batch_rewards))
                 step_thresholds.append(threshold)
             return step_rewards, step_thresholds
```

```
In [14]: step_rewards, step_threshold = run(percentile=50, policy=policy)
```

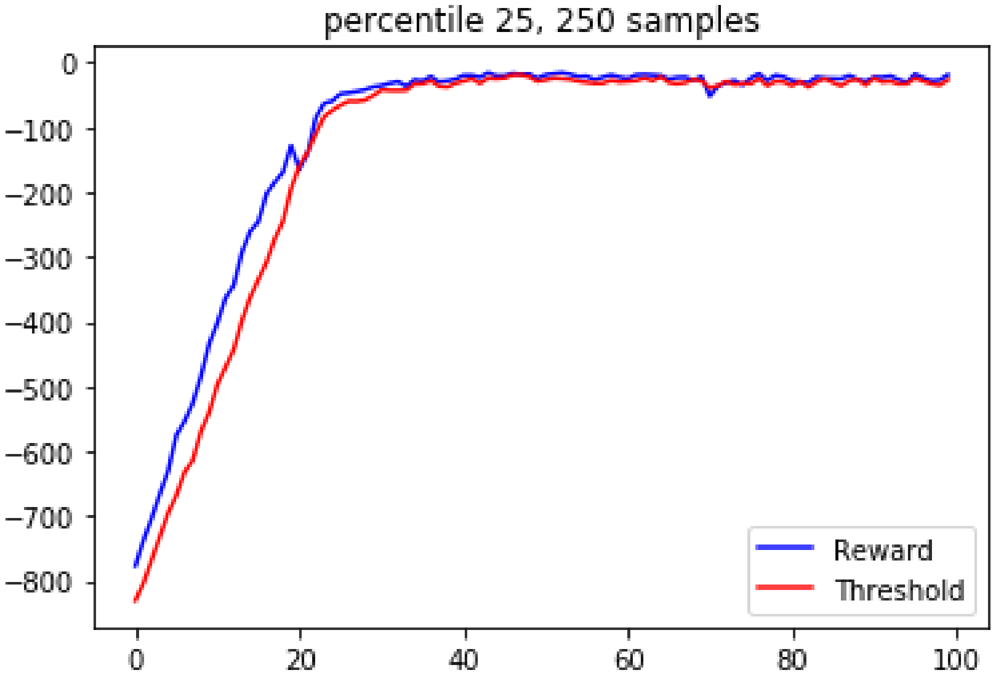
```
mean reward = -748.90000        threshold = -776.0
mean reward = -658.96800        threshold = -713.0
mean reward = -550.54000        threshold = -609.5
mean reward = -440.40000        threshold = -460.0
mean reward = -291.70400        threshold = -253.0
mean reward = -177.12400        threshold = -151.0
mean reward = -122.66000        threshold = -103.0
mean reward = -123.18400        threshold = -82.5
mean reward = -102.96400        threshold = -67.0
mean reward = -76.77200        threshold = -49.0
mean reward = -56.55600        threshold = -35.5
mean reward = -52.18000        threshold = -28.0
mean reward = -56.34400        threshold = -25.0
mean reward = -49.17600        threshold = -17.5
mean reward = -48.92400        threshold = -14.0
mean reward = -53.55600        threshold = -13.0
mean reward = -65.53200        threshold = -16.0
mean reward = -41.55600        threshold = -10.0
mean reward = -58.99200        threshold = -14.5
mean reward = -44.85600        threshold = -11.0
mean reward = -57.89600        threshold = -17.0
mean reward = -65.25200        threshold = -15.0
mean reward = -49.27200        threshold = -13.0
mean reward = -34.82000        threshold = -11.0
mean reward = -50.16800        threshold = -8.5
mean reward = -42.74400        threshold = -6.5
mean reward = -24.15200        threshold = -7.0
mean reward = -52.96800        threshold = -11.5
mean reward = -33.20000        threshold = -7.5
mean reward = -61.23600        threshold = -11.0
mean reward = -43.13200        threshold = -15.5
mean reward = -52.20000        threshold = -11.0
mean reward = -33.40800        threshold = -7.0
mean reward = -62.16800        threshold = -9.0
mean reward = -61.59600        threshold = -9.0
mean reward = -70.67600        threshold = -8.0
mean reward = -100.13600        threshold = -19.5
mean reward = -85.57600        threshold = -18.5
mean reward = -53.41600        threshold = -10.0
mean reward = -65.84000        threshold = -17.0
mean reward = -58.34000        threshold = -13.5
mean reward = -51.66400        threshold = -14.0
mean reward = -72.98400        threshold = -16.0
mean reward = -45.40400        threshold = -11.0
mean reward = -37.96400        threshold = -14.0
mean reward = -52.18400        threshold = -15.0
mean reward = -82.21600        threshold = -21.0
mean reward = -39.26800        threshold = -14.0
mean reward = -62.20800        threshold = -14.0
mean reward = -42.45200        threshold = -12.5
mean reward = -53.80000        threshold = -18.5
mean reward = -36.25200        threshold = -10.5
mean reward = -46.21200        threshold = -9.0
mean reward = -39.24800        threshold = -9.0
mean reward = -46.76000        threshold = -11.0
mean reward = -37.35600        threshold = -8.5
mean reward = -62.42000        threshold = -9.0
mean reward = -51.75200        threshold = -12.0
mean reward = -58.33600        threshold = -10.0
mean reward = -53.90800        threshold = -10.0
mean reward = -48.47600        threshold = -11.0
mean reward = -71.13600        threshold = -11.0
mean reward = -106.10800        threshold = -13.5
mean reward = -66.13200        threshold = -6.0
mean reward = -117.84800        threshold = -13.0
mean reward = -95.20800        threshold = -21.0
mean reward = -80.80800        threshold = -14.0
mean reward = -112.08400        threshold = -19.0
mean reward = -120.04800        threshold = -17.5
mean reward = -102.84800        threshold = -12.0
mean reward = -149.37600        threshold = -24.5
mean reward = -171.23200        threshold = -26.5
mean reward = -131.64800        threshold = -18.0
mean reward = -160.63200        threshold = -21.5
mean reward = -132.88800        threshold = -19.5
mean reward = -132.77600        threshold = -15.0
mean reward = -133.00800        threshold = -16.0
mean reward = -195.14800        threshold = -23.0
mean reward = -164.01200        threshold = -24.0
mean reward = -148.64400        threshold = -20.0
mean reward = -142.76800        threshold = -17.0
mean reward = -107.10400        threshold = -12.0
mean reward = -126.42000        threshold = -14.0
mean reward = -144.23200        threshold = -14.0
mean reward = -167.03600        threshold = -17.0
mean reward = -138.96400        threshold = -18.0
mean reward = -135.47200        threshold = -12.0
mean reward = -147.62800        threshold = -15.5
mean reward = -149.15600        threshold = -15.0
mean reward = -110.00400        threshold = -8.0
mean reward = -133.30000        threshold = -17.5
mean reward = -93.62800        threshold = -16.0
mean reward = -132.52000        threshold = -15.0
mean reward = -122.37600        threshold = -18.5
mean reward = -166.14400        threshold = -19.5
mean reward = -143.66400        threshold = -23.0
mean reward = -197.74000        threshold = -20.5
mean reward = -184.03600        threshold = -15.0
mean reward = -220.70400        threshold = -20.0
mean reward = -195.82800        threshold = -21.0
```

```
In [16]: plt.cla()
         plt.title("percentile 50, 250 samples")
         plt.plot(range(100), step_rewards, label='Reward', color="blue")
         plt.plot(range(100), step_threshold, label='Threshold', color="red")
         plt.legend()
         plt.show()
```



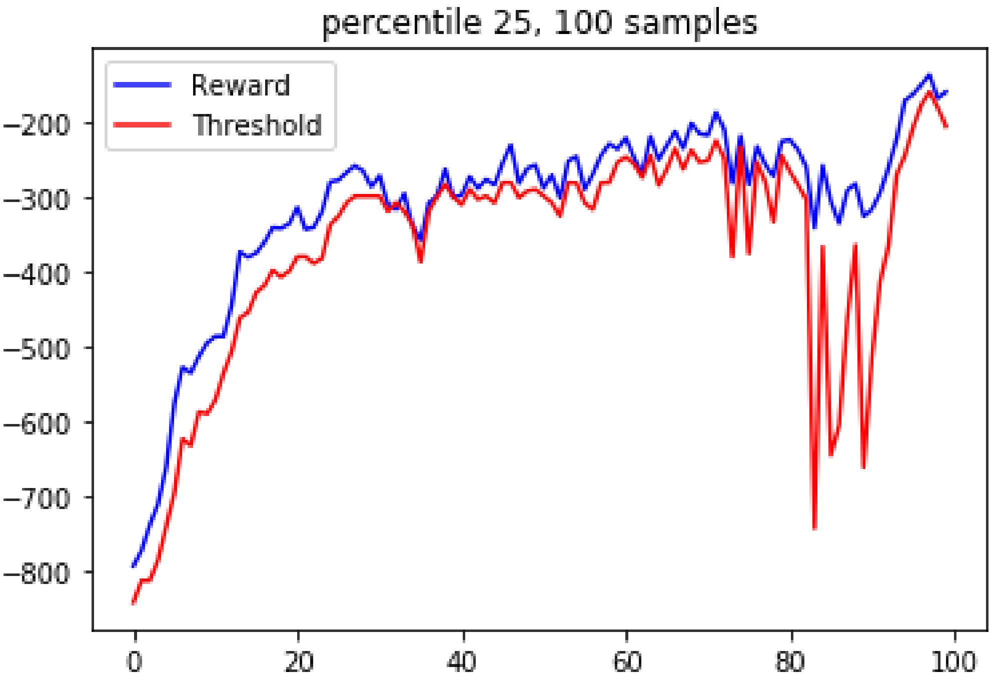
```
In [17]: step_rewards, step_threshold = run(percentile=25, policy=policy)
         plt.cla()
         plt.title("percentile 25, 250 samples")
         plt.plot(range(100), step_rewards, label='Reward', color="blue")
         plt.plot(range(100), step_threshold, label='Threshold', color="red")
         plt.legend()
         plt.show()
```

```
mean reward = -776.73600        threshold = -830.0
mean reward = -736.77600        threshold = -803.0
mean reward = -702.81600        threshold = -767.0
mean reward = -665.22000        threshold = -731.0
mean reward = -630.09200        threshold = -695.0
mean reward = -574.86000        threshold = -668.0
mean reward = -553.41200        threshold = -632.0
mean reward = -525.62400        threshold = -614.0
mean reward = -484.17600        threshold = -569.0
mean reward = -433.99200        threshold = -539.8
mean reward = -401.89200        threshold = -497.0
mean reward = -363.05200        threshold = -470.0
mean reward = -343.67600        threshold = -443.0
mean reward = -292.75200        threshold = -398.0
mean reward = -260.38000        threshold = -362.0
mean reward = -245.77600        threshold = -335.0
mean reward = -202.20800        threshold = -308.0
mean reward = -184.28000        threshold = -272.0
mean reward = -169.36800        threshold = -245.0
mean reward = -128.24000        threshold = -194.0
mean reward = -164.85200        threshold = -159.2
mean reward = -140.38800        threshold = -138.8
mean reward = -85.07600        threshold = -110.0
mean reward = -63.08000        threshold = -83.8
mean reward = -59.51200        threshold = -74.0
mean reward = -48.29200        threshold = -66.0
mean reward = -46.84800        threshold = -59.8
mean reward = -44.60000        threshold = -60.0
mean reward = -41.93600        threshold = -58.0
mean reward = -37.46800        threshold = -51.0
mean reward = -35.52800        threshold = -42.0
mean reward = -31.96000        threshold = -43.0
mean reward = -29.33200        threshold = -43.0
mean reward = -36.48400        threshold = -42.8
mean reward = -26.60400        threshold = -33.0
mean reward = -28.00000        threshold = -32.8
mean reward = -21.55600        threshold = -28.0
mean reward = -29.66000        threshold = -36.5
mean reward = -28.21200        threshold = -37.8
mean reward = -25.54400        threshold = -32.0
mean reward = -19.84800        threshold = -28.5
mean reward = -19.79200        threshold = -25.0
mean reward = -22.70000        threshold = -32.0
mean reward = -15.64800        threshold = -24.0
mean reward = -20.51600        threshold = -25.0
mean reward = -20.46400        threshold = -23.8
mean reward = -16.61600        threshold = -19.0
mean reward = -18.07600        threshold = -18.8
mean reward = -17.65600        threshold = -22.0
mean reward = -24.44400        threshold = -29.0
mean reward = -18.94400        threshold = -26.8
mean reward = -17.08400        threshold = -24.0
mean reward = -15.60400        threshold = -24.8
mean reward = -18.03600        threshold = -26.0
mean reward = -21.77200        threshold = -28.8
mean reward = -20.99600        threshold = -30.5
mean reward = -26.45200        threshold = -32.0
mean reward = -22.40000        threshold = -32.0
mean reward = -19.08000        threshold = -27.2
mean reward = -22.44000        threshold = -30.0
mean reward = -24.70800        threshold = -29.0
mean reward = -19.07600        threshold = -28.0
mean reward = -18.47200        threshold = -24.0
mean reward = -19.20400        threshold = -28.0
mean reward = -20.08400        threshold = -22.0
mean reward = -25.38000        threshold = -30.8
mean reward = -23.06800        threshold = -32.8
mean reward = -22.64800        threshold = -31.0
mean reward = -27.01600        threshold = -26.0
mean reward = -21.74400        threshold = -28.8
mean reward = -52.07200        threshold = -39.0
mean reward = -36.02400        threshold = -35.8
mean reward = -30.87200        threshold = -29.0
mean reward = -27.51200        threshold = -34.0
mean reward = -34.70400        threshold = -32.2
mean reward = -23.82000        threshold = -32.8
mean reward = -17.20800        threshold = -24.8
mean reward = -27.67200        threshold = -35.8
mean reward = -19.59200        threshold = -29.0
mean reward = -22.08800        threshold = -29.8
mean reward = -27.78400        threshold = -35.8
mean reward = -29.80000        threshold = -29.0
mean reward = -31.17200        threshold = -37.8
mean reward = -22.27200        threshold = -30.0
mean reward = -24.40400        threshold = -24.8
mean reward = -24.61600        threshold = -30.8
mean reward = -25.05600        threshold = -36.0
mean reward = -20.10800        threshold = -28.0
mean reward = -25.71200        threshold = -26.0
mean reward = -28.91200        threshold = -34.8
mean reward = -22.56800        threshold = -24.0
mean reward = -22.82800        threshold = -30.0
mean reward = -20.84800        threshold = -28.8
mean reward = -26.17600        threshold = -33.0
mean reward = -30.53600        threshold = -32.8
mean reward = -17.78000        threshold = -24.0
mean reward = -23.69200        threshold = -28.8
mean reward = -27.98800        threshold = -32.8
mean reward = -26.91200        threshold = -35.0
mean reward = -18.88400        threshold = -26.8
```



```
In [18]: step_rewards, step_threshold = run(percentile=25, n_samples=100, policy=policy)
         plt.cla()
         plt.title("percentile 25, 100 samples")
         plt.plot(range(100), step_rewards, label='Reward', color="blue")
         plt.plot(range(100), step_threshold, label='Threshold', color="red")
         plt.legend()
         plt.show()
```

```
mean reward = -792.57000        threshold = -841.2
mean reward = -771.99000        threshold = -812.0
mean reward = -737.87000        threshold = -812.0
mean reward = -710.15000        threshold = -785.0
mean reward = -660.84000        threshold = -740.0
mean reward = -575.75000        threshold = -695.0
mean reward = -527.63000        threshold = -623.0
mean reward = -535.13000        threshold = -632.0
mean reward = -512.77000        threshold = -587.0
mean reward = -494.94000        threshold = -589.2
mean reward = -486.33000        threshold = -571.2
mean reward = -486.10000        threshold = -535.2
mean reward = -444.22000        threshold = -506.0
mean reward = -373.74000        threshold = -461.2
mean reward = -380.68000        threshold = -454.2
mean reward = -374.88000        threshold = -427.2
mean reward = -360.46000        threshold = -418.2
mean reward = -341.11000        threshold = -398.0
mean reward = -342.03000        threshold = -407.0
mean reward = -336.11000        threshold = -399.5
mean reward = -314.08000        threshold = -380.0
mean reward = -343.51000        threshold = -380.0
mean reward = -340.60000        threshold = -389.0
mean reward = -319.88000        threshold = -382.2
mean reward = -280.08000        threshold = -337.2
mean reward = -277.83000        threshold = -326.0
mean reward = -267.47000        threshold = -308.0
mean reward = -258.79000        threshold = -299.0
mean reward = -265.33000        threshold = -299.0
mean reward = -286.38000        threshold = -299.0
mean reward = -271.62000        threshold = -299.0
mean reward = -309.74000        threshold = -319.2
mean reward = -316.79000        threshold = -308.0
mean reward = -295.78000        threshold = -319.2
mean reward = -338.03000        threshold = -337.2
mean reward = -357.66000        threshold = -386.8
mean reward = -308.34000        threshold = -319.2
mean reward = -298.45000        threshold = -299.8
mean reward = -263.14000        threshold = -283.2
mean reward = -300.87000        threshold = -300.8
mean reward = -297.79000        threshold = -310.2
mean reward = -273.05000        threshold = -290.0
mean reward = -288.52000        threshold = -303.5
mean reward = -276.76000        threshold = -299.0
mean reward = -284.06000        threshold = -308.0
mean reward = -255.02000        threshold = -281.0
mean reward = -230.85000        threshold = -281.0
mean reward = -281.84000        threshold = -301.2
mean reward = -262.25000        threshold = -292.2
mean reward = -257.86000        threshold = -290.0
mean reward = -288.43000        threshold = -299.0
mean reward = -270.75000        threshold = -308.0
mean reward = -301.73000        threshold = -326.0
mean reward = -252.31000        threshold = -281.0
mean reward = -246.31000        threshold = -281.5
mean reward = -290.06000        threshold = -308.0
mean reward = -269.62000        threshold = -317.0
mean reward = -244.89000        threshold = -281.0
mean reward = -229.35000        threshold = -281.0
mean reward = -236.55000        threshold = -254.0
mean reward = -221.31000        threshold = -247.2
mean reward = -248.10000        threshold = -256.2
mean reward = -269.00000        threshold = -274.2
mean reward = -219.51000        threshold = -245.0
mean reward = -250.57000        threshold = -284.5
mean reward = -231.63000        threshold = -263.0
mean reward = -212.85000        threshold = -236.0
mean reward = -235.22000        threshold = -263.0
mean reward = -201.96000        threshold = -238.0
mean reward = -215.89000        threshold = -254.0
mean reward = -217.81000        threshold = -251.8
mean reward = -187.37000        threshold = -225.5
mean reward = -210.01000        threshold = -248.8
mean reward = -281.58000        threshold = -380.2
mean reward = -218.85000        threshold = -233.2
mean reward = -283.76000        threshold = -375.8
mean reward = -233.10000        threshold = -254.0
mean reward = -257.05000        threshold = -280.2
mean reward = -272.42000        threshold = -333.5
mean reward = -226.27000        threshold = -245.2
mean reward = -223.83000        threshold = -266.2
mean reward = -237.65000        threshold = -284.2
mean reward = -260.44000        threshold = -303.5
mean reward = -341.36000        threshold = -742.2
mean reward = -258.53000        threshold = -366.2
mean reward = -306.26000        threshold = -644.8
mean reward = -335.42000        threshold = -605.0
mean reward = -291.34000        threshold = -463.2
mean reward = -282.35000        threshold = -364.2
mean reward = -326.48000        threshold = -661.2
mean reward = -316.78000        threshold = -515.0
mean reward = -294.85000        threshold = -413.8
mean reward = -261.13000        threshold = -365.2
mean reward = -222.61000        threshold = -270.2
mean reward = -171.61000        threshold = -245.0
mean reward = -163.07000        threshold = -207.5
mean reward = -150.65000        threshold = -177.2
mean reward = -137.08000        threshold = -160.0
mean reward = -168.80000        threshold = -181.0
mean reward = -159.95000        threshold = -206.0
```



## 6 Homework

### 6.0.1 Tabular correntropy method

You may have noticed that the taxi problem quickly converges from -10k to aroung -500 score (+- 500) and stays there. This is in part because taxi-v2 has some hard-coded randomness in the environment. Other reason is that the percentile was chosen poorly.

### 6.0.2 Tasks

- **1.1** (5 pt) Modify the tabular CEM (CrossEntropyMethod) code to plot distribution of rewards and threshold on each tick.
- **1.2** (5 pts) Find out how the algorithm performance changes if you change different percentile and different n_samples.

As expected, a smaller (but not too small) percentile converges slower but better better and fewer n_samples work worse.

- **1.3** (10 pts) Tune the algorithm to end up with positive average score.
- **1.4 bonus** (10 pt) Try to achieve a distribution where 25% or more samples score above +9.0

It's okay to modify the existing code.